TNO-report
TM-96-A053

TNO Human Factors
Research Institute

Kampweg 5
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

Phone +31 346 35 62 11
Fax +31 346 35 39 77

title

Automatic speech recognition performance in a simulation-based fast-jet cockpit application

authors
H.J.M. Steeneken
J.J. Kriekaard
D.A. van Leeuwen

date
28 November 1996

number of pages : 17 (incl. appendices, excl. distribution list)

TNO Human Factors Research Institute
is part of TNO Defence research
to which also belong:
TNO Physics and Electronics Laboratory
TNO Prins Maurits Laboratory

Netherlands Organization for
Applied Scientific Research

Managementuittreksel TNO Technische Menskunde, Soesterberg

titel : Automatische spraakherkenning toegepast voor controletaken in de cockpit van een jachtvliegtuig
auteurs : Dr.ing. H.J.M. Steeneken, J.J. Kriekaard en dr.ir. D.A. van Leeuwen
datum : 28 november 1996
opdrachtnr. : A91/KLu/325
IWP-nr. : 786.4
rapportnr. : TM-96-A053

De toepassing van automatische spraakherkenning voor het besturen van controlefuncties in de cockpit van een jachtvliegtuig werd bestudeerd door TNO Technische Menskunde en het Nationaal Lucht- en Ruimtevaartlaboratorium. Het project omvatte drie fasen: (1) onderzoek naar de haalbaarheid, (2) keuze van een commerciële herkenner en inbouw in een vluchtsimulator, (3) evaluatie tijdens realistische simulatorvluchten met jachtvliegers als proefpersoon.
Recentelijk zijn de experimenten voor fase 3 afgerond. De resultaten worden in dit rapport gepresenteerd. Er werden meer dan 20 sorties uitgevoerd (testvluchten van ca. een uur). De resultaten van 17 sorties, die werden uitgevoerd door drie ervaren vliegers, werden geanalyseerd. De experimenten werden uitgevoerd in de F-16 simulator van het NLR. De vliegers konden middels spraak de instelling van de radio, displays en HOTAS functies bedienen (hands on throttle and stick). Deze systemen konden ook manueel worden bediend conform de gebruikelijke bedrijfssituatie. Gedurende de "vluchten" werd zowel het spraaksignaal opgenomen als een video opname gemaakt van de handelingen van de vlieger.
Analyse van de acties van de vlieger alsmede de spraakbesturing en de debriefing werden uitgevoerd door het NLR en worden separaat gerapporteerd.
In dit rapport wordt de prestatie van de herkenner geanalyseerd. Er werd vastgesteld dat onder de gegeven simulatorcondities de prestaties van de herkenner afnemen van 0,95 (accuracy) voor gelezen spraak tot 0,69 voor de spontane spraak conditie zoals deze optreedt in de simulator.
De resultaten van vier, elders uitgevoerde, experimenten sluiten hierbij aan. Ook hier werden scores boven 0,9 gevonden voor herkennerexperimenten op basis van gelezen commando's en scores van ca. 0,7 voor spontane spraak.
Van de vocabulaire van 281 woorden werden slechts 65 woorden frequent gebruikt (met 90% dekking). Dit betekent dat indien de herkenners met deze kleinere vocabulaire worden toegepast een betere herkenning wordt verkregen. Van al de registraties van de spraaksignalen werd een databestand gemaakt. Bij dit databestand zijn al/ de uitspraken orthografisch geannoteerd (de tekst is per commando beschikbaar). De beschrijving van dit databestand is in een apart rapport beschikbaar.
Met dit databestand werd een experiment uitgevoerd met een moderne foneem/grammatica gebaseerde herkenner. Deze herkenner was sprekeronafhankelijk en getraind voor Amerikaans/Engels. Hierbij werd van breedbandige spraaksignalen gebruik gemaakt (niet via zuurstofmasker en zonder lawaai). De behaalde gemiddelde accuracy bedroeg 0,85. Voor de drie sprekers was dat resp.: 0,90, 0,90 en 0,74. Het ligt in de verwachting dat met dit type herkenner, getraind met representatieve spraaksignalen (zuurstofmasker, sprekers die niet hun moedertaal spreken), een score boven de 0,95 mogelijk is. Experimenten met dit type herkenner zullen in de nabije toekomst door ons worden uitgevoerd.

CONTENTS Page

SUMMARY 3

SAMENVATTING 4

1 INTRODUCTION 5

2 LABORATORY EVALUATION 5
2.1 Results 6
2.2 Discussion 8

3 SIMULATOR EVALUATION 8
3.1 NLR real-time results 9
3.2 Replay in laboratory 9
3.3 Analysis of the results 10
3.4 Discussion 13

4 MAJOR FINDINGS 15

REFERENCES 17

Report No.:	TM-96-A053
Title:	Automatic speech recognition performance in a simulation-based fast-jet cockpit application
Authors:	Dr.ing. H.J.M. Steeneken, J.J. Kriekaard, Dr.ir. D.A. van Leeuwen
Institute:	TNO Human Factors Research Institute Group: Perception
Date:	November 1996
DO Assignment No.:	A91/KLu/325
No. in Program of Work:	786.4

SUMMARY

A project on automatic speech recognition for control of systems in a fast-jet cockpit was conducted by the TNO Human Factors Research Institute (TNO-HFRI) and the National Aerospace Laboratory (NLR). The project included three phases: (1) the feasibility, (2) speech recognizer selection and implementation in a flight simulator and (3) performance testing in an advanced fast jet simulator.
Presently the experiments for phase 3 were conducted, the results are given in this report. The experiments consisted of over 29 sorties of approximately one hour each. In total the results of 17 sorties, performed by three experienced pilots, were analyzed. During each sortie a pilot in the F-16 National Simulator Facility had access to a control by voice of radio systems, displays and HOTAS functions (hands-on-throttle-and-stick). These systems could also be controlled manually as in the normal situation. During the "flight" tests recordings were made of the speech signals and a video recording of the pilot actions.
Analysis of all pilot actions including the voice control and debriefing was performed by the NLR and is reported separately. In this report the recognizer performance is analyzed. It was found that under these simulator flight conditions the performance (accuracy) drops from over 0.95 for read speech to 0.69 for the simulator spontaneous speech condition. Results obtained in four flight experiments performed in other laboratories showed similar results for read speech (three experiments) and for spontaneous speech (one experiment).
From the original 281 word vocabulary only 65 words were used frequently by the pilots. These 65 words had a coverage of 90% of all words used during the tests. This means that the complexity of the recognition process can be reduced which will lead to a better performance of the recognizer.
From the speech material a calibrated date base was built with all the speech utterances annotated orthographically at command string level. This data base is described in a separate report.
A pilot study was performed with a modern phoneme/grammar based recognizer. With this speaker independent system a mean performance of 0.85 (accuracy) was obtained. It is expected that this performance will exceed the 0.95 if this type of recognizer is trained for the non-native English speaking pilots rather than for, probably read, American English speech. Also training with more representative speech signals obtained through an oxygen mask is required. It is foreseen that we will perform experiments with such a system in the near future.

Rap.nr. TM-96-A053

TNO Technische Menskunde
Soesterberg

Automatische spraakherkenning toegepast voor controletaken in de cockpit van een jachtvliegtuig

H.J.M. Steeneken, J.J. Kriekaard en D.A. van Leeuwen

SAMENVATTING

De toepassing van automatische spraakherkenning voor het besturen van controlefuncties in de cockpit van een jachtvliegtuig werd bestudeerd door TNO Technische Menskunde en het Nationaal Lucht- en Ruimtevaartlaboratorium. Het project omvatte drie fasen: (1) onderzoek naar de haalbaarheid, (2) keuze van een commerciële herkenner en inbouw in een vluchtsimulator, (3) evaluatie tijdens realistische simulatorvluchten met jachtvliegers als proefpersoon.
Recentelijk zijn de experimenten voor fase 3 afgerond. De resultaten worden in dit rapport gepresenteerd. Er werden meer dan 20 sorties uitgevoerd (testvluchten van ca. een uur). De resultaten van 17 sorties, die werden uitgevoerd door drie ervaren vliegers, werden geanalyseerd. De experimenten werden uitgevoerd in de F-16 simulator van het NLR. De vliegers kon middels spraak de instelling van de radio, displays en HOTAS functies bedienen (hands on throttle and stick). Deze systemen konden ook manueel worden bedient conform de gebruikelijke bedrijfssituatie. Gedurende de "vluchten" werden zowel het spraaksignaal opgenomen als een video opname gemaakt van de handelingen van de vlieger.
Analyse van de acties van de vlieger alsmede de spraakbesturing en de debriefing werd uitgevoerd door het NLR en worden separaat gerapporteerd.
In dit rapport wordt de prestatie van de herkenner geanalyseerd. Er werd vastgesteld dat onder de gegeven simulatorcondities de prestaties van de herkenner afnemen van 0,95 (accuracy) voor gelezen spraak tot 0,69 voor de spontane spraak conditie zoals deze optreedt in de simulator.
De resultaten van vier, elders uitgevoerde, experimenten sluiten hierbij aan. Ook hier werden scores boven 0,9 gevonden voor herkennerexperimenten op basis van gelezen commando's en scores van ca. 0,7 voor spontane spraak.
Van de vocabulaire van 281 woorden werden slechts 65 woorden frequent gebruikt (met 90% dekking). Dit betekent dat indien de herkenners met deze kleinere vocabulaire worden toegepast een betere herkenning wordt verkregen. Van al de registraties van de spraaksignalen werd een databestand gemaakt. Bij dit databestand zijn all de uitspraken orthografisch geannoteerd (de tekst is per commando beschikbaar). De beschrijving van dit databestand is in een apart rapport beschikbaar.
Met dit databestand werd een experiment uitgevoerd met een moderne foneem/grammatica gebaseerde herkenner. Deze herkenner was sprekeronafhankelijk en getraind voor Amerikaans/Engels. Hierbij werd van breedbandige spraaksignalen gebruik gemaakt (niet via zuurstofmasker en zonder lawaai). De behaalde gemiddelde accuracy bedroeg 0,85. Voor de drie sprekers was dit resp.: 0,90, 0,90 en 0,74. Het ligt in de verwachting dat met dit type herkenner, getraind met representatieve spraaksignalen (zuurstofmasker, sprekers die niet hun moedertaal spreken), een score boven de 0,95 mogelijk is. Experimenten met dit type herkenner zullen in de nabije toekomst door ons worden uitgevoerd.

1 INTRODUCTION

In 1991 a project on automatic speech recognition for control of systems in a fast-jet cockpit was started. The project included three phases: (1) the feasibility, (2) speech recognizer selection and implementation in a flight simulator, and (3) performance testing in an advanced simulator (originally in an aircraft). Presently the third phase of the project is finished, the results are given in various reports of which this report concerns the performance of automatic speech recognition in a flight simulator during "realistic" sorties. Other reports concern the simulator environment and pilot responses, and the performance of present recognition technology as the recognizer used in this test had to be selected a few years ago. These present tests were performed at the NLR in the National Simulator Facility (NSF) in a time period of two months. Experienced (test) pilots participated in the experiments. During the test flights real-time recognition was obtained and selected functions were controlled. The control included HOTAS functions (hands on throttle and stick), radar and radio controls, and weaponry functions. A total of 281 control words combined in a syntax structure were defined. During the test the speech signals were recorded for later evaluation and for the collection of a representative data base. The report concerning the simulator tests is prepared by the NLR (Pijpers & Eertink, 1996).
In this report the laboratory optimalization of the recognizer used for the experiments (Marconi MR-8) is described. Also an analysis of the test results as obtained with the simulator experiment and repetition of the recognition test in the laboratory with the original speech data is described.

2 LABORATORY EVALUATION

Prior to the simulator test the selected recognition system was evaluated in various representative conditions, simulated in the laboratory. A major part of this study was already performed during phase II of the project. For example the effect of the use of an oxygen mask, speech level variation and a high environmental noise level on the recognition performance was studied and reported by Steeneken and Kriekaard (1995a,b).
In the present laboratory evaluation two additional aspects were studied (1) the use of representative speakers (pilots), and (2) the optimal parameter setting of the recognizer in combination with the signal treatment as studied in phase II. It should be stated that studies make use of speech data-base based on read speech items. The use of spontaneous speech can only be obtained under more realistic conditions as was scheduled for the simulator experiments in phase III.
In general the optimal tuning and the effect of setting recognition parameters is not given by the vendors of commercial recognition systems. Trial and error methods are normally proposed. Therefore, a systematic study on the effect of the relevant parameter settings was performed. As the tests require much effort, a computer controlled test-bed was used in which the parameter setting, training and test material setting could be predefined in a control file. The following parameters were varied:

- speakers,
- training parameters of the recognizer,
- recognition parameters of the recognizer.

The performance is expressed by an accuracy figure which is based on the number of correct responses, inserted words (insertions and false alarms), and the total number of spoken words according:

$$\text{Accuracy} = \frac{\text{words correct} - \text{insertions}}{\text{total}} \qquad (1)$$

This commonly used accuracy measure was also used in the earlier studies (Steeneken & Kriekaard, 1995a).

2.1 Results

The major part of the recognition algorithm is predefined and cannot be adjusted by the user. Only the acceptance threshold (a measure of the fit between template and presented speech item) and five parameters related to utterance detection can be set. For the recognition of connected words the detection of word boundaries is an important task of the recognizer and related to the signal quality in a particular application. With the Marconi MR-8 five parameters are related to start and end-point detection. Parameters 'Alim' and 'Blim' define the energy threshold for begin and end detection, 'Fore' and 'Aft' define the additional number of frames to be taken into account before and after the threshold trigger, and 'Nlim' defines the number of frames that the Alim threshold must be exceeded before a trigger is generated.
The optimum setting of these parameters may depend on the speaker, the status of the recognizer (training or testing), noise corruption and acoustical aspects of the speech signal. The parameter values can typically be varied in 8−15 steps. This results into such a large number of possible combinations that it is not feasible to investigate all these conditions. In a controlled experiment the parameter settings were changed individually while the other parameter values were kept constant at a default value. In this way the relation between system performance and parameter value was obtained. It was found that for most parameters the parameter setting did not have a major effect on the recognition performance. For the parameter Nlim a strong effect on the performance was found.
The evaluation experiment was performed with a vocabulary of 75 test words selected from the cockpit vocabulary and without a syntax. Two speakers were used and two noise conditions (no noise and representative cockpit noise). The speech was recorded by making use of the oxygen mask and the AGC-amplifier (automatic gain control). A maximum accuracy over 90% was obtained. Separate experiments were performed for the parameter setting for training and test.

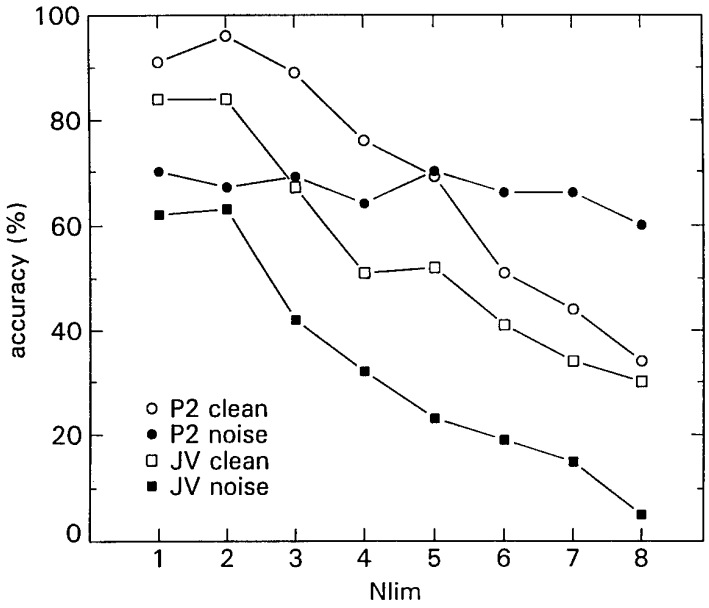


Fig. 1 Recognition accuracy as a function of the relative parameter setting of 'Nlim' for four conditions (2 speakers, 2 noise conditions).

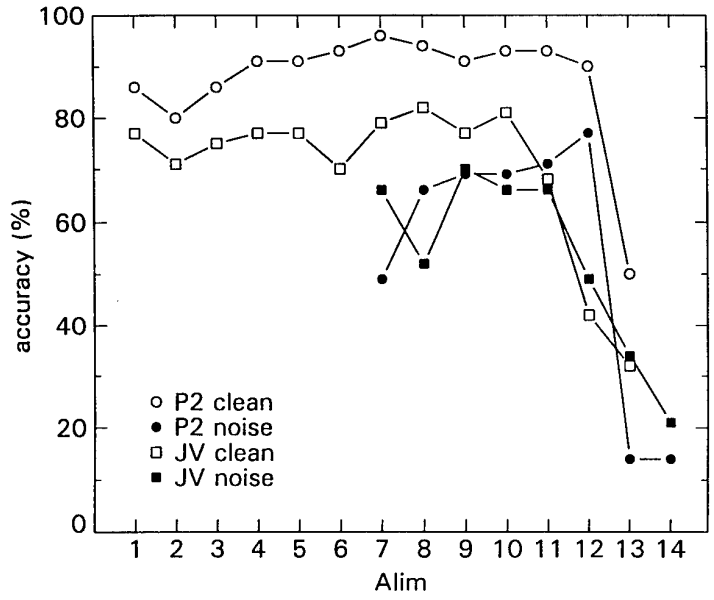


Fig. 2 Recognition accuracy as a function of the relative parameter setting of 'Alim' for four conditions (2 speakers, 2 noise conditions).

In Fig. 1 the recognition performance (accuracy) is given as a function of the relative parameter value for parameter Nlim in a test session. The relation is given for four conditions: two speakers and two noise conditions. A relative parameter value of 1−3 offers an optimum for all combinations of speaker and noise. The performance as a function of the parameter value of Alim is less dependent of the parameter setting. A fairly flat function is obtained as shown in Fig. 2. This was also the case for the other parameters (Alim, Blim, Fore, Aft) in conditions of both training and testing.

The "optimal" values based on these results for the test in the NSF are:

parameter	training	testing
Alim	7	10
Blim	3	5
Fore	5	8
Aft	8	10
Nlim	2	1

2.2 Discussion

Recognition performance depends primary on the algorithm chosen for the recognition process. This is predefined with the recognizer selected for this project. This recognizer is based on a front-end analysis of the speech signal with the mel-cepstrum method and a matching algorithm between templates and test item by a dynamic time warping algorithm (DTW). Additional to this fixed signal treatment, the training procedure and the setting of specific parameters for the recognition algorithm can be set by the user. In a previous phase of the project the specific signal processing (oxygen mask, AGC-amplifier) was developed and evaluated.
In the present laboratory evaluation the optimal setting of five recognition parameters was determined. It was found that the parameter setting only slightly depends on the speaker or combination with background noise of the speech signal. One parameter showed a clear optimum (Nlim) for both the training and the test condition. As the optimal parameter settings were very close to the settings used before we advised the NLR to use these settings for the planned experiments in the NSF.

3 SIMULATOR EVALUATION

The goal of the project on voice control of cockpit systems is to perform a realistic experiment and to obtain the subjective pilot response and objective performance measures. During the project it was decided to perform this experiment in the National Simulation Facility for the MLU-F16 at the NLR. The selected recognition system including the signal processing features was installed in the NSF and an interface to relevant controls (display, radio, weaponry and HOTAS) was established. During realistic preprogrammed flights (sorties), of 70 min average duration, the pilot could operate a number of systems either by voice or manually. A detailed report on these sorties is given by Pijpers and Eertink (1996).
In total 29 sorties were performed. Within this set 17 sorties were performed by three highly trained pilots and the results of these sorties were used for the evaluation. The total flight time amounts to 18 hrs. The total set of speech utterances for the control task amounts to 134 min. A real-time recognition and control of systems was performed. The voice control task was organized in separate nodes related to specific functions (radio, display, etc.). A total of 281 words were available. A detailed description of the syntax is given by Pijpers and Van Zutphen (1995). Three pilots participated in the final experiments. These pilots

were familiar with the MLU-F16 cockpit. From the real-time experiments a performance measure of the recognition was obtained. The speech signals were also recorded for later analysis and to repeat the experiment under laboratory conditions. This is relevant for conditions where syntax errors, false PTT-triggers (push-to-talk), hesitations, etc. define the performance of the presently used recognizer.
In the next sections the results of the real-time performance, the laboratory replay and an error analysis are given.

3.1 NLR real-time results

During the experiments an estimation of the performance was produced by monitoring the spoken command and recognizer responses. This led to an estimated score which was used at the debriefing of the pilot. Sometimes when a low performance was obtained or when certain commands did not work properly the training of the recognizer was extended with an additional training session. This led to the condition that the training may vary during the course of an experiment with a particular pilot.
During the tests the output of the recognizer was stored into a logfile together with the speech signal and the PTT actions. In the laboratory the speech signal was transcribed (annotated) orthographically. Hence, from all spoken utterances the written version was available. After time alignment of the recognizer response (logfile) and the annotation file the performance can be obtained.
An automatic scoring program gives the correct, deleted and inserted words. From this the accuracy according to Eq. (1) can be obtained. The accuracy measure thus obtained is given in Table I (header column NSF). The accuracy is very speaker dependent. Speaker 2 gives the highest scores (acc. 0.81). The mean of all three speakers is 0.69. Errors may be introduced by a poor recognition performance of the system but also by the speakers (e.g., syntax errors, pronouncing out-of-vocabulary words (OOV's), and incorrect PTT-actions). An analysis on the structure of the errors is given in § 3.3.

3.2 Replay in laboratory

As all the speech material was transcribed to computer files, annotated, and corrected for operation errors of the pilots a robust data base was obtained. This data base is described in a separate report (Steeneken *et al.*, 1996b). The data base was used to repeat the simulator experiment in the laboratory automatic recognizer test-bed with the MR-8 recognizer and with other recognizers. All settings of the MR-8 recognition chain were identical to the settings used during the simulator experiments. The templates used for the training were those with the latest update. In general a higher score (approx. 0.04) was obtained with the test-bed evaluation because control errors of the pilot had been eliminated.
For the experiments a strict syntax for the commands to the system was used. This means that only a part of the vocabulary is active at any given time. The mean perplexity is approximately 13.5. Average effective vocabulary size at any instant of time with the same data base the test was also repeated without making use of a syntax. As the pilots did not use all the 281 words of the vocabulary a reduced set of words was used. With 65 words

90% of the total material is covered, with a perplexity of 65, which in theory is a more difficult recognition task. The results for these laboratory tests are also given in Table I.

Table I Description and performance of the 17 sorties performed in the NSF.

sortie	pilot	total length (min)	annotated length (min)	NSF	word accuracy lab replay	% no syntax, 65 words
1	**2**	82.0	7.9	0.81	0.78	0.78
2		82.0	7.8	0.83	0.82	0.80
3		67.2	6.5	0.78	0.90	0.80
4		82.0	8.8	0.78	0.83	0.84
5		70.0	7.7	0.69	0.90	0.89
6		70.3	6.4	0.88	0.93	0.86
7		75.0	5.9	0.87	0.94	0.85
			mean speaker **2**	0.81	0.87	0.83
8	**3**	67.0	11.3	0.61	0.80	0.75
9		43.0	8.3	0.63	0.69	0.70
10		40.0	6.6	0.55	0.55	0.74
11		68.3	12.3	0.67	0.64	0.73
12		68.3	11.5	0.74	0.79	0.65
13		75.1	10.5	0.78	0.69	0.62
			mean speaker **3**	0.66	0.69	0.70
14	**4**	41.0	4.5	0.63	0.61	0.70
15		60.0	7.0	0.53	0.59	0.70
16		53.7	7.2	0.65	0.65	0.74
17		60.0	5.6	0.60	0.72	0.69
			mean speaker **4**	0.60	0.64	0.71
			mean **all** speakers	0.69	0.73	0.76

3.3 Analysis of the results

A statistical analysis of the speech items used during the sorties was performed. The 17 annotated sorties amount a total of 18 hrs from which 134 min of speech utterances were detected. The data base consists of 12231 words within 5825 utterances. In total 175 different words were used, however 42 of these words were not included within the original vocabulary of 281 words (OOV, out-of-vocabulary word). In Fig. 3 the cumulative frequency distribution of all words used is given based on the total data base (solid line). Also an analysis was performed on the individual pilot data. These cumulative frequency distributions are also given in Fig. 3 (dotted lines). The word sequence for the individual pilots were the same as the sequence for the total analysis. Comparison of the individual results with the total results shows that only a slightly different use of the vocabulary was employed by the three pilots.

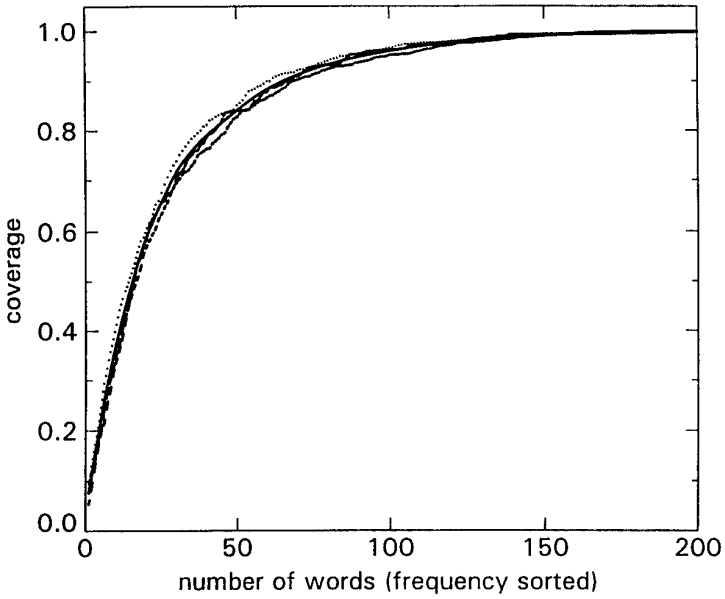


Fig. 3 Cumulative distribution of the word frequency from the 281 word vocabulary and the OOV words (solid line). The individual pilot distributions (same word sequence) are also given (dotted lines).

The figure also shows that with the use of only 66 words (including one OOV-word) a coverage of 90% of all words is obtained.

The scoring of the results as presented in Table I was performed automatically with dedicated scoring software. Also a scoring was performed by hand for the results of three sorties, one for each pilot respectively sortie number 3, 12, 16 according to Table I. An elaborated analysis on the type of errors was performed. The total number of utterances in the three selected sorties was 1158 (2428 words). From the recognition results of these utterances a number of 110 errors could be traced to circumstances for which the recognizer cannot be held responsible. The reason for these errors can be the wrong use of the syntax or words, or of technical nature. The type of errors that were identified and selectively corrected are given in Table II.

Table II

number of words	type of error	corrected accuracy (accumulated)	remarks
2428	all data	0,732	original data
38	not in node	0.756	(e.g., *switching v_h_f// select*)
28	out of syntax	0,77	(e.g., *master v_h_f*)
22	out of vocab.	0,781	(e.g., *jammer stand_by*)
14	out of vocab.	0,787	OOV but correctly recognized (e.g., *give me* rather than *gimme*)
8	other	0,792	PTT errors, radio speech, etc.

Most of these errors were due to incorrect commands of the speakers who used words which were not valid at a certain state of the syntax node or not within the vocabulary. Excluding the all analysis errors of the three sorties leads to an increase of the accuracy. This is given in Table III. The mean accuracy for investigated sorties increased from 0.72 to 0.79. Pilot 3 and 4 (who have in general a lower score than pilot 2) show the largest improvement for correction of speaking errors.

Table III Comparison of the accuracy of three sorties where a correction for speaking errors is applied.

sortie	pilot	NSF uncorrected	NSF corrected
3	2	0.78	0.81
12	3	0.74	0.81
16	4	0.65	0.73
	mean	0.72	0.78

For the assessment of the recognition procedure, analysis at word level is appropriate in order to detect insertions, deletions and misclassifications expressed in the word error rate or the accuracy. However, for the control task it is relevant to consider the performance at command level, expressed by the utterance error rate. It was found that there is a high correlation between error rates at word and utterance level (correlation coefficient $r=0.96$).

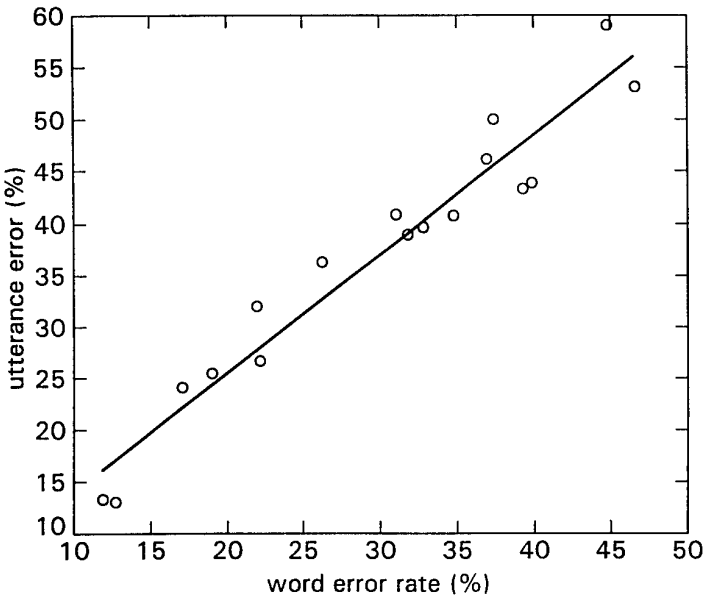


Fig. 4 Relation between utterance and word error rate for the 17 investigated sorties.

In Fig. 4 the relation between these error rates is given for the 17 sorties. The best fitting linear dependence is described by:

$$\text{Utterance error rate} = 1.2 \times \text{word error rate} + 2\% \quad (r=0.96)$$

As the regression of 1.2 is very close to one and the high correlation coefficient it can be concluded that the errors cluster within an utterance.

3.4 Discussion

The performance of the MR-8 recognizer during the simulator experiments gives a mean accuracy of 0.69. The performance is much lower than the performance obtained during the initial laboratory experiments with the same vocabulary where an accuracy over 0.95 was obtained. The use of the recognizer during realistic simulator experiments had a major influence and the mean accuracy dropped from above 0.95 to 0.69. Analysis indicated that this decrease can be explained by two effects (1) reduction of the speech quality during operations in a multi task scenario, and (2) speaking and syntax errors. It was also shown that the influence on the recognition performance of these factors is reduced for more modern grammar based recognition systems.
The main difference between the two experiments is that for the laboratory test *read* speech was used while in the simulator *spontaneous* speech was obtained. The present results are closer to a realistic flight condition than the earlier laboratory tests. The results show to be pilot dependent. For pilot 4 the mean accuracy is 0.60 while pilot 2 produces a mean accuracy of 0.81. An acceptable system accuracy should be higher than 0.95 (South, 1996). The present recognizer does not offer this performance in combination with the adverse speech input quality. Quality aspects such as speech produced inside the oxygen mask and environmental noise can be dealt with, but additional aspects as spontaneous speech, hesitations, inaccurate pronunciations and the complex syntax structure reduce the performance.

The mean perplexity (number of words active at a certain syntax node) was 13.5, this resulted into a mean recognition accuracy of 0.69. Analysis of the words used by the pilots indicated that a 90% coverage could be obtained with only 65 words of the 281 word vocabulary. Replay of the simulator experiment with the same speech tokens show an improvement of the accuracy from 0.69 to 0.76, although the perplexity is much higher in the latter condition (harder task for the recognizer). This is in fact an unusual behaviour due to the "unknown" rejection criterion in the MR8 when the system is used with a complex syntax structure (Hunt, private communication).

As reported by Eertink and Pijpers the pilots had difficulties with the complex command structure as is required for the MR8. The syntax was rather complex consisting of 281 words and was also very strict. Over 300 nodes allowing for 4000 node-to-node connections is too much. The pilots were sometimes unaware of the node-status of the recognizer. This subject is discussed in the report on the flight experiments by Eertink and Pijpers (1996). More recent recognition systems allow a more flexible structure of the word strings. These

systems are based on a statistical language model which can be trained and hence are domain specific. For the application discussed here cockpit control commands obtained from the limited vocabulary are suitable to train such a system.

Few errors were made by incorrect use of the push-to-talk switch. From a human factors point of view no PTT actions are recommended. A word spotting facility or an open mike would be preferred. The present technology would probably generate too many errors for a reliable operation.

Modern ASR systems

It was foreseen that during the life-time of the project the state-of-the-art of automatic speech recognition would improve. At the moment the decision had to be made the selection of the MR-8 system was an optimal choice for integration in the NSF. For this reason all the speech data have been recorded, annotated, calibrated, and made available on a CD-ROM. With this data base ASR performance measurements can be repeated with more recent systems without making use of the NSF.

Present day recognizers are speaker independent and can handle a large vocabulary. Language models offer a flexible recognition of a command even if an error in the command structure is included. A rigid syntax is not required while this was a primary request for the recognizer used in the present experiment (i.e., the MR-8).

We performed an evaluation of such a system and found that a performance (accuracy) of 0.90, 0.90, and 0.74 could be obtained for pilots 2, 3 and 4 respectively (see Steeneken *et al.,* 1996b). It should be noticed that this system was trained for American-English (not by the pilots) with speech signals with a better quality than achieved with the oxygen mask microphone. The grammar was adapted to the command strings as used in the cockpit. For this purpose the command strings for half of the sorties (8 sorties) were used to build a language model, while the speech signals of the other half, based on a different set of command strings, has been used to perform the test.

Other studies on ASR in flight applications

Application of voice control is presently studied in France, Germany, UK, and USA. All these studies make use of connected word recognizers similar to the system used in this study. In three out of four countries flight experiments were performed in which no real control of the cockpit systems was included as was done in the experiments described above. In summary the following results were obtained:

France (Cordonnier *et al.,* 1996). In this study flight tests were performed with two types of command strings: setting the radio frequency with connected digits and control of a display engine. There was real control of the radio system but the display control was artificial. The speech data were recorded for later evaluation with a specially designed connected word recognizer. The performance of the system increased during separate tests of the speaker

from a rate of 89% correct to 98% correct. The test was performed with speakers who were highly trained in performing recognition experiments.

Germany (Prévôt & Onken, 1995). A connected speech recognizer (MR8) was used to control an on-board pilot assistance system. The system was evaluated during simulator and real flight experiments. Two pilots were involved for both the recognizer performance (percentage correct commands) improved during the tests from approx 63% to 86%. The experiments were performed in a standard aircraft, hence no oxygen mask was used. The vocabulary size is not given.

UK (South, 1996). The experiments performed at Farnborough included flight experiments with 4g turns and centrifuge experiments at various g-levels. The speakers were supplied with a pressure breathing system similar to the system used in a jet aircraft. The test consisted of connected digits which were read by the speakers. The recognizer used for the experiments was a Marconi ASR-1000 which is a similar systems as the MR-8. For g-force levels up to 3g an average accuracy was obtained between 0.9 and 0.95. For g-force of 4g and up to 6g the performance dropped from 0.86 to 0.65. The results also indicated that for an experienced aircrew and runs up to 4g a smaller effect of g-force on the recognition performance was found than for other speakers.

USA (Williamson, 1996). An ITT connected word, speaker dependent recognizer was used for the recognition experiment with speech samples recorded during flights. For this purpose a 54 word vocabulary was used. Twelve pilots participated in the tests. The experiment was based on read speech recorded through a boom microphone (M-162) in an aircraft on the ground and during flights with g-force of 1g and 3g. The noise level was approximately 115 dB SPL. There was no equipment controlled by the recognizer. The recognition performance (accuracy) was at 1g 0.977 and at 3g 0.971.

The results of the various experiments summarized above show that for the vocabulary sizes between 20 an 54 words an accuracy of 0.95 can be obtained even at g-forces up to 3g. All these experiments were performed with either read speech or spontaneous speech utterances consisting of digit strings. No spontaneous speech samples uttered during the performance of other tasks as flying an aircraft or control of equipment has been used. Also in all experiments a similar type of recognizer was used. The performance of the MR-8 results reported here is similar to the results obtained in the other studies.

4 MAJOR FINDINGS

The recognition performance during (simulator) flight conditions (mean accuracy 0.69) is much lower than during laboratory tests (accuracy $>$ 0.95). The main reason is that the speech quality produced during a primary flight task is much lower than during reading well defined commands such as used for training the recognizer. The pilots made many syntax errors, hesitations, etc. The recognition system used for the experiments (connected word recognizer) is not configured for handling this type of errors.

Results obtained in four flight experiments performed elsewhere showed similar results for read speech (three experiments) and for spontaneous speech (one experiment).

The syntax was too complex for the pilots. They only used 65 words from the original 281 word vocabulary. These 65 words had a coverage of 90% of all words used during the tests. The pilots were not always aware of the node status of the recognizer which resulted in rejection of the command. For a future system a simpler syntax is therefore required.

In total 17 sorties of approximately one hour each were considered in the results reported here. Three pilots participated.

From the speech material a calibrated date base was built with all the speech utterances annotated orthographically at command string level. This data base is described in a separate report.

A pilot study with a modern phoneme/grammar based speaker independent recognizer showed a mean accuracy of 0.85. It is expected that this performance will exceed the 0.95 if this type of recognizer is trained for the non-native English speaking pilots rather than for read American English. Also training with more representative speech signals obtained through an oxygen mask is required. Unfortunately such a modern recognition system which can be trained by a user was not yet available. It is expected however that we will perform experiments with such a system in the near future.

REFERENCES

Cordonnier, A., Kientz, D. & Wassner, H. (1996). Commande Vocale sur Aeronefs Militaires: Evaluation en Vol et Resultats. *Proc. AGARD Symposium "Audio effectiveness in aviation",* Copenhagen (in press).

Eertink, B.J. & Pijpers, E.W. (1996). *Evaluation of integrated automatic speech recognition on the NSF Mid-life update F-16 simulator* (Report NLR CR 96646 L **III.4** Volume I and II). Amsterdam: National Aerospace Laboratory (NLR).

Houtgast, T. & Van Velden, J.G. (1992). *Application of automatic speech recognition in the fighter cockpit. Summary report on phase I* (Report IZF 1992 A-24, DMKLu/AC02/A/9105-report **I.0**). Soesterberg: TNO Institute for Perception.

Pijpers, E.W. & Van Zutphen, W.J.C.M. (1995). *Development and simulator implementation of an automatic speech recognition application for the mid-life update F-16 Cockpit* (Report NLR CR 95220 L **II.3**). Amsterdam: National Aerospace Laboratory (NLR).

Prévôt, T. & Onken, R. (1995). In-flight evaluation of CASSY: A system providing intelligent on-board pilot assistance. *Air Traffic Control Quarterly, 3 (3),* 183-204.

South, A. (1996). Effect of Aircraft Performance on Voice Recognition systems. *Proc. AGARD Symposium "Audio effectiveness in aviation",* Copenhagen (in press).

Steeneken, H.J.M. (1995). *Development and performance of a cockpit control system operated by voice; summary report on phase II (DMKLu/AC02/A/9105-Report II.0)* (Report TNO-TM 1995 A-42 **II.0**). Soesterberg: TNO Human Factors Research Institute.

Steeneken, H.J.M. (1996). *Development and performance of a cockpit control system operated by voice; summary report of project DMKLu/AC02-A/9105* (Report TM-96-A055 **III.0**). Soesterberg: TNO Human Factors Research Institute.

Steeneken, H.J.M. & Kriekaard, J.J. (1995a). *Development and evaluation of the electro-acoustical input environment for an automatic speech recognizer in cockpit applications* (Report TNO-TM 1995 A-40 **II.1**). Soesterberg: TNO Human Factors Research Institute.

Steeneken, H.J.M. & Kriekaard, J.J. (1995b). *Prediction of the performance of automatic speech recognition for control applications in a fast-jet cockpit* (Report TNO-TM 1995 A-41 **II.2**). Soesterberg: TNO Human Factors Research Institute.

Steeneken, H.J.M., Kriekaard, J.J. & Van Leeuwen, D.A. (1996a). *Automatic speech recognition performance in a simulation-based fast-jet cockpit application* (Report TM-96-A053 **III.1**). Soesterberg: TNO Human Factors Research Institute.

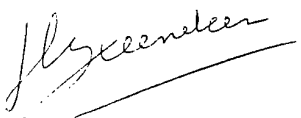
Steeneken, H.J.M., Kriekaard, J.J. & Van Leeuwen, D.A. (1996b). *Spontaneous-speech data base for cockpit control applications applied to commercial state-of-the-art speech recognition technology* (Report TM-96-A054 **III.2**). Soesterberg: TNO Human Factors Research Institute.

Steeneken, H.J.M. & Pijpers, E.W. (1996). Development and Performance of a cockpit control system operated by voice. *Proc. AGARD Symposium "Audio effectiveness in aviation",* Copenhagen (in press).

Steeneken, H.J.M. & Van Velden, J.G. (1989). RAMOS—Recognizer assessment by means of manipulation and speech. *European Speech Conf. ESCA,* Paris.

Williamson, D.T. (1996). Flight Test Performance Optimization of ITT VRS-1290 Speech Recognition System. *Proc. AGARD Symposium "Audio effectiveness in aviation",* Copenhagen (in press).

Soesterberg, 28 November 1996



Dr.ing. H.J.M. Steeneken
(1st author, project manager)

REPORT DOCUMENTATION PAGE

1. DEFENCE REPORT NUMBER (MOD-NL)	2. RECIPIENT'S ACCESSION NUMBER	3. PERFORMING ORGANIZATION REPORT NUMBER
RP 96-0196		TM-96-A053
4. PROJECT/TASK/WORK UNIT NO.	5. CONTRACT NUMBER	6. REPORT DATE
786.4	A91/KLu/325	28 November 1996
7. NUMBER OF PAGES	8. NUMBER OF REFERENCES	9. TYPE OF REPORT AND DATES COVERED
17	15	Interim

10. TITLE AND SUBTITLE

Automatic speech recognition performance in a simulation-based fast-jet cockpit application

11. AUTHOR(S)

H.J.M. Steeneken, J.J. Kriekaard and D.A. van Leeuwen

12. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO Human Factors Research Institute
Kampweg 5
3769 DE SOESTERBERG

13. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director of Airforce Research and Development
Binckhorstlaan 135
2516 BA DEN HAAG

14. SUPPLEMENTARY NOTES

15. ABSTRACT (MAXIMUM 200 WORDS, 1044 BYTE)

A project on automatic speech recognition for control of systems in a fast-jet cockpit was conducted by the TNO Human Factors Research Institute (TNO-HFRI) and the National Aerospace Laboratory (NLR). The project comprised performance testing in an advanced fast jet simulator.

In total the results of 17 sorties, performed by three experienced pilots, are presented. During each sortie the pilot had access to a control by voice of radio systems, displays and HOTAS functions. During the "flight" tests recordings were made of the speech signals and a video recording of the pilot actions.

Analysis of all pilot actions including the voice control and debriefing was performed by the NLR and is reported separately. In this reported the recognizer performance is analyzed. It was found that under these simulator flight conditions the performance (accuracy) drops from over 0.95 for read speech to 0.69 for the simulator spontaneous speech condition. Results obtained in four flight experiments performed in other laboratories showed similar results for read speech (three experiments) and for spontaneous speech (one experiment).

From the original 281 word vocabulary only 65 words were used frequently by the pilots. These 65 words had a coverage of 90% of all words used during the tests. This means that the complexity of the recognition process can be reduced, which will lead to a better performance of the recognizer.

From the speech material a calibrated date base was built with all the speech utterances annotated orthographically at command string level. This data base is described in a separate report.

A pilot study was performed with a modern phoneme/grammar based recognizer. With this speaker independent system a mean performance of 0.85 (accuracy) was obtained. It is expected that this performance will exceed the 0.95 if this type of recognizer is trained for the non-native English speaking pilots. Also training with more representative speech signals obtained through an oxygen mask is required.

16. DESCRIPTORS	IDENTIFIERS
Automatic Speech Recognition Cockpit Command & Control	

17a. SECURITY CLASSIFICATION (OF REPORT)	17b. SECURITY CLASSIFICATION (OF PAGE)	17c. SECURITY CLASSIFICATION (OF ABSTRACT)
18. DISTRIBUTION/AVAILABILITY STATEMENT Unlimited availability		17d. SECURITY CLASSIFICATION (OF TITLES)

VERZENDLIJST

1. Directeur M&P DO

2. Directie Wetenschappelijk Onderzoek en Ontwikkeling Defensie

3. { Hoofd Wetenschappelijk Onderzoek KL
   Plv. Hoofd Wetenschappelijk Onderzoek KL

4. Hoofd Wetenschappelijk Onderzoek KLu

5. { Hoofd Wetenschappelijk Onderzoek KM
   Plv. Hoofd Wetenschappelijk Onderzoek KM

6, 7 en 8. Bibliotheek KMA, Breda

9 t/m 13. Maj.ir. J.H.J. Verhulst, Koninklijke Luchtmacht, Afdeling Wetenschappelijke Ondersteuning, DM/MXS, Den Haag

Extra exemplaren van dit rapport kunnen worden aangevraagd door tussenkomst van de HWOs of de DWOO.